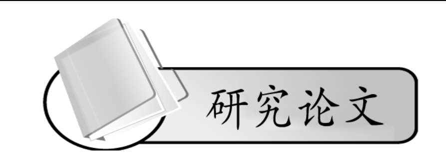


改进潜在语义分析和支持向量机算法用于突发安全事件舆情预警*

田世海　吕德丽

(哈尔滨理工大学管理学院　哈尔滨 150040)

摘要:【目的】针对现有预警体系多以企业自身和监管部门为主体、忽视网络舆情，导致预警力度不强、缺乏透明度及敏感性、使突发性安全问题时有发生且无法得到及时处理的现状，提出一种新的舆情预警模型。【方法】通过元搜索技术挖掘舆情信息，增加基准偏移值优化情感特征项倾向性权重，添加修正因子以改进潜在语义分析和支持向量机(LSA+SVM)算法，构建舆情分类预警模型。【结果】以多组突发性安全事件为例，应用 Matlab 进行仿真实验。结果证明该舆情预警模型切实可行，反应迅速，在语义维度为 10 时准确率可达 85.75%。【局限】此方法对于能引起关注和讨论的安全事件更加有效。【结论】改进算法适用于舆情预警，可为企业和监管部门根据分类结果及时采取有效的预警措施提供合理化建议。

关键词: 潜在语义分析　支持向量机　舆情预警　情感倾向性分析

分类号: G203

1　引　言

网络舆情具有传播速度快、渠道多和范围广等特点，其针对热点事件和突发事件的传播、扩散以及发酵对于企业的决策及管理起到重要作用。然而，舆情信息纷繁杂乱，具有强烈的情感色彩和干扰噪声，甚至可能威胁到企业的生存与发展。因此，如何妥善利用网络舆情，对企业相关的舆情信息进行及时的分类预警并采取措施理应受到企业及学者的重点关注。

在舆情预警方面，国内外学者开展了大量的研究。吴鹏等[1]通过 Agent 建模，构建了网络群体行为模型。Li 等[2]以人工神经网络为对象预测产品产量安全。王兰成[3]分析了舆情情报的功能，设计了针对突发事件应急处置的舆情情报支援系统架构。Papetti 等[4]提出一个基于多因素和多数据源舆情的预警模型，通过多个案例进行验证，新的预警模型在减少预警时间和源数据的条件下，依然可以保证预警信息的准确性。董凯欣等[5]通过分析角色指标和子群挖掘意见领袖，对舆情机制提出建议。陈福集等[6]通过建立意见交互机制有效预测舆情事件发展趋势。

综上所述，大多数舆情预警研究是在整体层面对预警措施进行建模和预测，然而，情感特征词的分布较整个模型来说不够均衡，且语义维度复杂，因此需要从优化语义维度和速度的视角，深入研究更加精准的分类方法。本文通过改进潜在语义分析(Latent Semantic Analysis, LSA)和支持向量机(Support Vector Machine, SVM)算法建立舆情分类预警模型，提高倾向性预测的准确性，改善分类的效率及舆情状况感知，

通讯作者：吕德丽，ORCID：0000-2347-7112-1342，E-mail：lvdelixx@126.com。

*本文系国家自然科学基金项目“高技术虚拟产业集群运行模式研究”(项目编号：70873029)、黑龙江省自然科学基金项目“高新技术企业物流模式选择技术研究”(项目编号：G201203)和黑龙江省博士后科研启动资金资助项目“黑龙江省制造企业动态联盟信誉保障机制研究”(项目编号：LBH-Q12065)的研究成果之一。

以确保企业在风险进一步扩大之前采取积极有效措施，同时根据舆情反馈进一步解决自身问题，创新产品，适应市场要求。

2 舆情分类预警模型构建

网络舆情除了在传播上具有巨大优势之外，还包含以下 4 点较特殊的性质:

(1) 受国家政策法规影响较大。国家在控制、检验和管理等各个方面的安全标准都随着安全事件的发生及技术的进步实时更新，这是企业预警需要着重考虑的因素之一。

(2) 突发性较强，具有缓时性，发酵时间较长。安全事件往往由突发事件引起，并以极快的速度传播，吸引大量的关注度。安全事件往往会牵涉到企业的生产管理制度、行业的检验制度等，会长时间地传播、发酵和沉淀。

(3) 受众关注度广，覆盖面较强。由于网络内容与日常生活息息相关且涉及到每个个体的安全，大众倾向于投入更多的关注，直至事件解决。

(4) 对企业的影响及打击较大。突发性安全事件对于企业的打击往往是致命的，如"霸王致癌"风波、"三聚氰胺"事件，因此企业应倾注更多资源在危机预警上。

根据以上特质，舆情分类预警模型需要尽可能降低舆情倾向性的语义维度，以便企业能够在安全事件发生的初级阶段迅速地捕捉舆情，还需要优异的组合和分类能力，及时对事件评级并准确判定其倾向性。LSA 可以在文本分析中消除同义词和多义词造成的偏差，获得更准确的文本向量，同时简化文本向量，提高计算效率；SVM 作为泛化能力优异的分类器被广为应用，并能够推广应用到函数拟合等其他机器学习问题中[7-9]。因此本文选择 LSA 和 SVM 算法组合来满足分类预警的要求，并加以适当的改进，使其更符合舆情主体的特征。舆情分类预警模型的构建主要包括以下几个步骤：首先进行舆情分类预警流程分析；其次确定和修正情感特征词权重，改进 LSA+SVM 算法；最后进行算法模型的实现。

2.1 舆情分类预警流程

舆情分类预警流程主要分为信息抓取、倾向性判定和舆情分类三个环节。信息抓取利用元搜索技术和 Nutch 爬虫，对抓取的数据进行简单的降噪、清洗及分词处理，提取情感特征词。本研究重点在于倾向性判定及分类。舆情分类预警流程如图 1 所示。

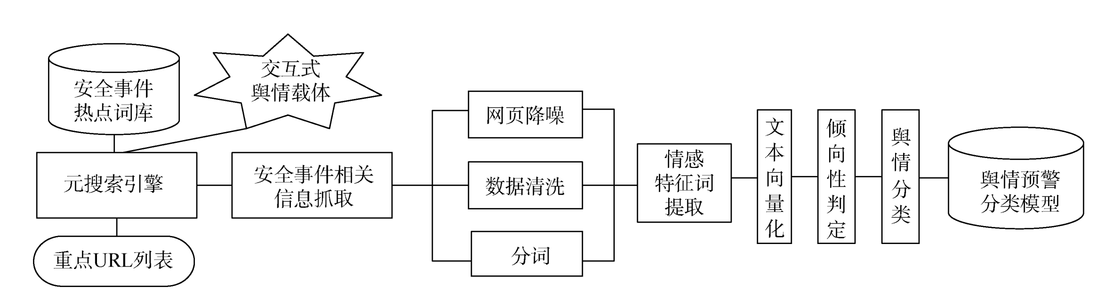


图 1 舆情分类预警流程

2.2 舆情分类情感特征词权重确定及改进

选取安全事件相关的热点关键词，利用开源工具 Nutch 爬虫进行舆情语料挖掘，得到舆情的相关 URL 列表。结合 HowNet 的情感分析词语集词典中的相关标准，提取情感倾向性特征词，以形容词、副词和名词为主，向量化情感特征项文本并按以下格式存储:

$$\left(a_i, T, link, t_i, r, W\right)$$

其中，a_i 代表舆情分类的情感特征项；T 代表获取特征项的时间；t_i 代表获取该特征项的相关文本发布时间；r 为二值型字段代表该 URL 是否被转载；W 表示该特征项来源网页的重要性权值。当 r 为"是"时，W 取该情感特征项的权值。考虑到信息来源的影响程度和信息的语义倾向性对相关企业造成的后果，在极大程度上会影响相关舆情特征项的权重。设出现在此舆情文本向量中情感特征项的重要程度为 $tfidf_{ik}$[10]。

$$tfidf_{ik} = tf_{ik} \times idf_k$$

其中，tf_{ik} 表示情感特征项 a_i 出现频率。

$$tf_{ik} = \frac{n_{ik}}{N_i}$$

其中，n_{ik} 表示特征项 a_i 出现次数，tf_{ik} 需要结合

在整个文本向量中出现的特征项总量 N_i 来计算。

idf_k 表示该情感特征项 a_i 的逆文档频率，即在整篇文章中出现较少但特征明显存在的词汇，因此要计算特征项 a_i 数目的倒数，选取该值的对数来计算：

$$idf_k = \log \frac{N}{n_k}$$

然而在现实的舆情文本中，长句中副词、名词涉及较多，情感倾向性更为明显，导致特征项的权重值对长文本更加偏袒，造成log函数为零，失去对判断的影响。同时，安全事件中普遍用包含语意和语态的经验系数来突出重要特征项，国家政策和法规的临时发布或改进会对相关行业造成显著影响。由此，为解决log函数为零的问题将其值增加0.01，为临时法规政策的发布及时改进主权重值，添加基准偏移值 *offset* [11]，从而得到舆情分类情感特征项权重值求解公式为：

$$W = \frac{tf(t_k, d_j) \cdot \log(N / df(t_k) + 0.01)}{\sqrt{\sum_{P=1}^{K} \left[tf(t_k, d_j) \cdot \log(N / df(t_k) + 0.01) \right]^2}} \times offset$$

通过舆情分类情感特征项权重公式对向量化文本的权重值进行求解并储存，以便于下一步对向量空间化分类。

2.3　舆情分类情感特征词空间向量化及分类

舆情分类的情感特征词以单一文本向量储存，不属于同一个概念空间，空间维度太高，需要降维，以便进行组合及分类。基于改进 LSA+SVM 算法的情感特征词分类方法基本流程如图 2 所示。

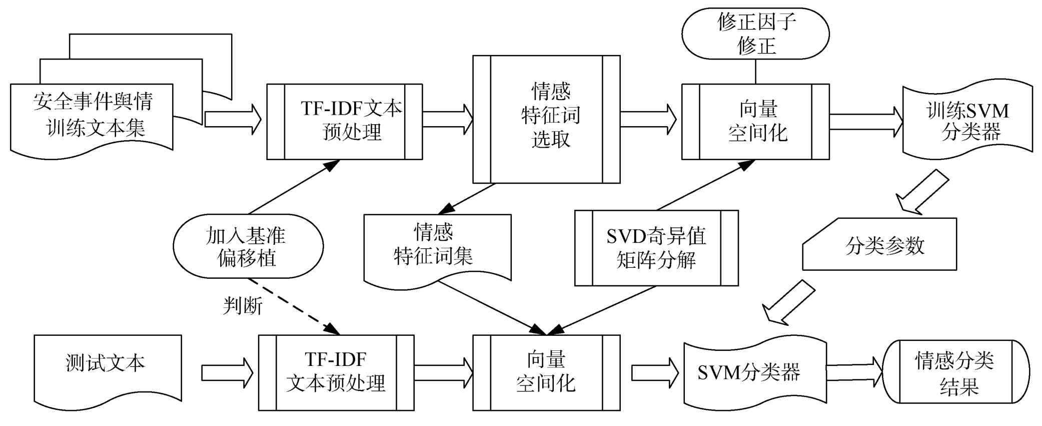


图 2　舆情分类情感特征词分类流程

其中，文本预处理即权重值计算和改进的过程。LSA通过奇异值分解将舆情分类文本分割成不同局部特征空间，避免了一词多意和一意多词等噪声项的干扰，使舆情情感特征词所表达的含义更加明确且更易被感知。分解已经向量化的舆情特征词空间向量，即以 $m \times n$ 的矩阵格式储存：

$$\boldsymbol{A} = (\mathrm{a}_{ij})_{m \times n}$$

对情感特征项矩阵进行初步处理，如果多个舆情情感特征词属于同义词，语义相关度较高，则将其划分为同一类别；相对来说，不同类别的特征词出现同义的概率就会较低。由此，将矩阵 $\boldsymbol{A}$ 分解为多个不同类别的矩阵集合的组合形态，如下：

$$\boldsymbol{A} = \boldsymbol{USV}^T$$

其中，$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 都分别是 $\boldsymbol{A}^T\boldsymbol{A}$ 的左右奇异向量矩阵，那么 $\boldsymbol{S} = \{\beta_1, \beta_2, \cdots, \beta_r\}$ 为矩阵 $\boldsymbol{A}$ 的奇异值矩阵，满足 $\beta_1 \geqslant \beta_2 \geqslant \cdots \geqslant \beta_r \geqslant 0$。用奇异值分解(Singular Value Decompositon, SVD)对整个 $\boldsymbol{USV}^T$ 空间实施压缩处理，获得 k 秩矩阵，形式如下：

$$\boldsymbol{A}_k = \boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}_k^T$$

其具体分解过程如图 3 所示。$\boldsymbol{S}_k$ 表示分解后的基本奇异值矩阵并已按其语义相关性分解为多局部矩阵。

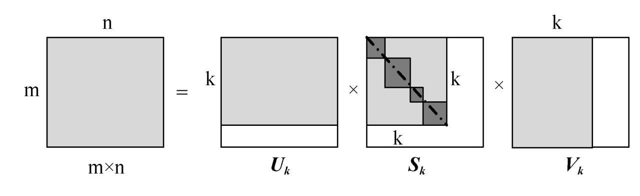


图 3　奇异值分解过程

特征词相似度关系通过矩阵 $\boldsymbol{A}_k$ 行向量之间的内积 $\boldsymbol{A}_k\boldsymbol{A}_k^T$ 来计算:

$$\boldsymbol{A}_k\boldsymbol{A}_k^T=(\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^T)(\boldsymbol{V}_k\boldsymbol{S}_k\boldsymbol{V}_k^T)=(\boldsymbol{U}_k\boldsymbol{S}_k^2\boldsymbol{V}_k^T)$$
$$=(\boldsymbol{U}_k\boldsymbol{S}_k)(\boldsymbol{U}_k\boldsymbol{S}_k)^T=\boldsymbol{SS}^T$$

计算得到的 $\boldsymbol{SS}^T$ 表示第 i 、j 行的内积关系, 反映出两个向量之间的异同, k 表示降维后的维数。得到新的文本向量, 将该文本向量送至 SVM 分类模块按照相关性分类。

2.4　舆情分类的 LSA+SVM 算法改进

突发性安全事件发酵期长、受众广以及对企业信誉影响较大, 普通的分类器难以判定其情感倾向和危险程度, 需要在其特征词局部矩阵中添加修正因子 O_{a_i} [12-13]。修正因子主要以在该局部特征向量中发现情感词 a_f 和程度副词 a_g 同时出现作为基准, 将其权重相乘, 所得值作为矩阵严重程度的优先判断标准。计算方法为:

$$O_{a_i}=W_{a_f}\times W_{a_g}$$

将 $\boldsymbol{A}_k=\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^T$ 中 $\boldsymbol{S}_k$ 以修正因子 O_{a_i} 和奇异值分解后得到的局部矩阵重新排列, 同时, 根据安全特征词的权重值, 加入基准偏移值, 使其原排列方式和趋势产生偏差。对几种奇异值的线性关系进行模拟, 模拟为一条具有相关性的回归跳跃曲线, 如图 4 所示。

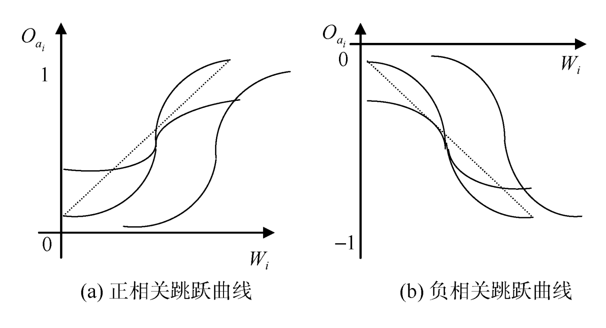


图 4　修正后的奇异值相关性跳跃曲线

当 $O_{a_i}>0$ 时, 相关性跳跃曲线如图4(a)所示, 该舆情分类局部矩阵的正面意义较强, 一般出现在对该事件持乐观态度甚至有助于企业品牌形象的评论或文章中。当 $O_{a_i}=0$ 时, 曲线无明显跳跃性, 一般在横轴附近波动, 则该局部矩阵的倾向性趋近于中立。此类评论更趋近于叙述事实, 不包含明显的批判或者支持行为。当 $O_{a_i}<0$ 时, 相关性跳跃曲线如图4(b)所示, 该局部矩阵的负面意义较强, 说明该文对该事件反应强烈, 对企业有明显的批判态度。其中 O_{a_i} 值越接近 1 或 −1, 则该特征词情感倾向性越严重。

2.5　改进 LSA+SVM 算法实现过程

(1) 训练算法实现

选取大量舆情分类训练文本对改进 LSA+SVM 算法进行训练, 形成标准的舆情分类预警参数模型, 得到 α,β,χ 即惩罚函数系数、线性最大间隙、核函数系数三个基本参数, 训练算法具体过程[14]如下。

输入: 特征词向量集合 $A=\{a_1,a_2,\cdots a_n\}$ 、基准偏移值 *offset*

输出: 分类参数模型 $M=\{\alpha,\beta,\chi\}$

For $i=1:m, j=1:n$

$\{tf_{i,j}=tfidf_{i,j}*offset$

$A=[tf_{i,j}]_{m*n}*O_{ai}$

$A_r=[U]_{m\times r}\times[S]_{r\times r}\times[V]_{r\times n}\xrightarrow{SVD}A_k=[U]_{m\times k}\times[S]_{k\times k}\times[V]_{k\times n}$

$X_k\xrightarrow{SVM}M=\{\alpha,\beta,\chi\}\}$

(2) 测试算法实现

测试算法结合参数模型和 SVM 分类器对新的特征项文本进行情感倾向性分类, 先根据修正因子的正负性分为两个层次, 再依靠权重划分成特重舆情(S 级)、重度舆情(A 级)、中度舆情(B 级)、轻度舆情(C 级)和需要关注(D 级)5 个等级, 将正向舆情纳入到企业反馈信息和创新信息中记作(P 级)[15]。测试算法过程如下。

输入: 待分类测试特征词集合 $A'=\{a_1,a_2,\cdots a_n\}$ 、基准偏移值 *offset*

输出: 分类结果 $Tab=\{S,A,B,C,D\}$

For $i=1:m, j=1:n$

$\{tf_{i,j}=tfidf_{i,j}*offset$

$A_t=[tf_{i,j}]_{m*n}*O_{ai}$

$A_r=[U]_{m\times r}\times[S]_{r\times r}\times[V]_{r\times n}\xrightarrow{SVD}A_k=[U]_{m\times k}\times[S]_{k\times k}\times[V]_{k\times n}$

$X_k\xrightarrow{SVM+\{\alpha,\beta,\chi\}}Tab=\{S,A,B,C,D,P\}\}$

根据基准偏移值和修正因子的修正以及大量文本训练, 可以让该模型更加准确高效地对实时安全事件实施危机情况进行判定, 并将判定结果及时反馈给企业, 达到预警目的。

3　舆情分类预警实现与仿真

为确保舆情分类更为准确, 以三类不同领域的突发性安全事件来讨论 LSA+SVM 算法的现实应用, 分

别是食品安全为代表的蒙牛黄曲霉素事件(事件一)、互联网用户安全为代表的百度“莆田系”事件(事件二)和生产安全为代表的天津滨海化工厂泄露事件(事件三)作为分析对象。首先根据舆情类别选取“热词+舆情词汇”格式，依据元搜索技术使用 Python 在各个搜索引擎热点新闻中设计爬虫，挖掘该系列字段，获取900 余篇事件相关文章及评论 URL 列表，如图 5 所示。

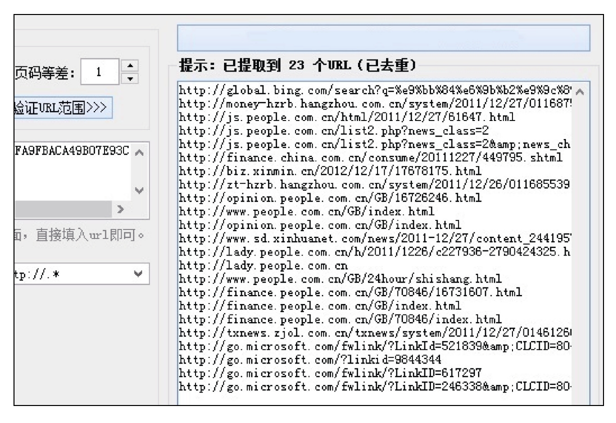


图 5　URL 获取列表

对抓取的文档进行简单的去重降噪，选取中国科学院计算技术研究所 ICTCLAS 分词系统以及 LibSVM 作为仿真软件，经权重计算及改进 LSA+SVM 算法，通过交叉验证来获取基础的分类参数模型。算法实现基于 Windows7 操作系统，仿真软件为 Matlab2012b，训练中选取径向基(Radius Basis Function，RBF)核函数利用交叉检验的方法来确定最优的参数模型及分类模型。

训练所得核函数系数约为 0.431，惩罚函数系数为 0.424462，得到倾向性为负向的有效特征向量共 324 个，倾向性为正向或中性的有效特征向量共 198 个。根据最终分类跳跃曲线的运动情况可发现跳跃曲线更加趋近于负相关，舆情倾向为负。对分类模型在不同的语义维度下实施对比实验，以在不同参数下文档倾向性的准确率作为衡量其性能的基本指标:

$$C = (P_P + N_N)/(P + N)$$

其中，P 代表选取的正面文档总数，P_P 代表选取时为正面文档且分类后 $O_{a_i} > 0$，仍为正面文档；类似地，N 表示选取时为负面文档，N_N 代表分类后仍为负面文档的文档数量。随机选取三组语义维度的不同取值，即 k 分别为 5, 10, 15 来进行准确率计算，结果如图 6 所示。

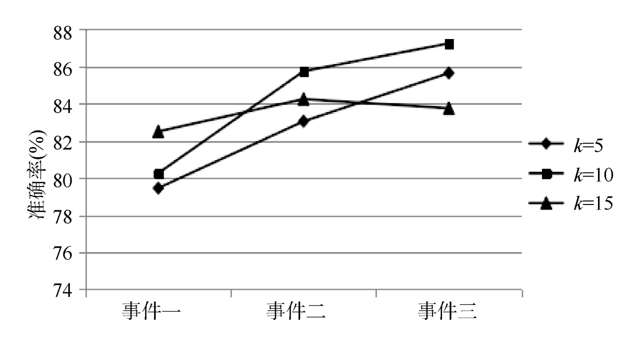


图 6　语义维度对分类结果准确率影响

可知，当 k=10 时表现最好，准确率可达 87.25%，可以高效体现出文本的相关特性。维度太低易导致结果偏差，而维度太高时易发生语义混乱导致分级不够准确[15]。分类算法实现结果如图 7 所示。

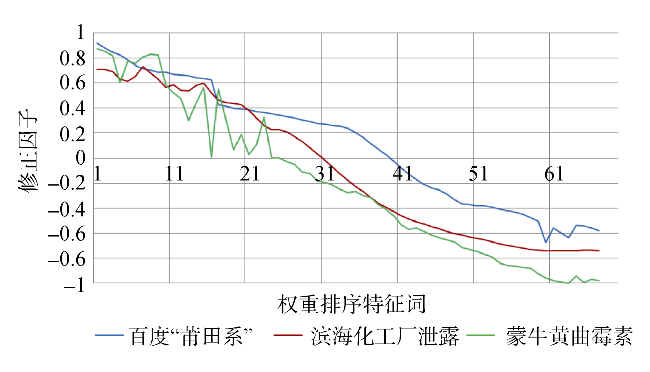


图 7　算法实现结果

将最终得到的有效文档及评论按照修正因子的值进行分类，正面舆情都可归为 P 级，负面舆情即需要引起警报的 S, A, B, C, D 这 5 个等级在(–1, 0)区间上等距离划分，由此得到 S∈[–1, –0.8]，A∈[–0.8, –0.6]，B∈[–0.6, –0.4]，C∈[–0.4, –0, 2]，D∈[–0.2.0)。但考虑到(0, 0.1)区间虽属正面舆情但倾向性不明显，所以将该区间划分到 D(需要关注)中[16-17]。除去重复性文档和训练文档，得到上述三个突发性安全事件正负面文档占比如表 1 所示。

表 1　三个突发性安全事件正负面文档比率

突发性安全事件	负面文档占比(%)	中性文档占比(%)	正面文档占比(%)
百度“莆田系”事件	59.8	6.7	33.5
滨海化工厂泄露事件	65.5	27.3	7.2
蒙牛黄曲霉素事件	76.3	9.8	13.9

通过正负文档的比率，可以初步看出不同突发性安全事件的舆情倾向性，在此选取负面文档及中性文档，进一步归纳不同等级的舆情文档数量，以判别各突发性安全事件的紧急程度，从而判断企业应采取何种措施。

等级分类及主要舆情词汇如表 2 所示。

表 2　三个突发性安全事件等级分类及主要舆情词汇

突发性安全事件	主要舆情词汇	舆情等级	区间	数量	是否含基准偏移值
百度"莆田系"事件	作恶；丑闻；互相勾结；虚假宣传；垂死挣扎；医疗伦理缺失；无底线；贪心；造假系；谋财害命；毒瘤；作孽；肮脏的广告手段；不道德；放纵；缺乏监督；不作为	S	[–1, –0.8)	31	Y
		A	[–0.8, –0.6)	29	N
		B	[–0.6, –0.4)	21	N
		C	[–0.4, –0.2)	13	N
		D	[–0.2, 0.1)	7	N
滨海化工厂泄露事件	毒害百姓；强烈抗议；生命财产得不到保护；气味刺鼻；恶心头晕；告状无门；不顾百姓死活；隐患巨大；污染；寝食难安；惨烈；极度危险；扼腕堵心；吸取教训	S	[–1, –0.8)	41	Y
		A	[–0.8, –0.6)	27	Y
		B	[–0.6, –0.4)	25	N
		C	[–0.4, –0.2)	8	N
		D	[–0.2, 0.1)	20	N
蒙牛黄曲霉素事件	无需怜悯；毫无原则；显然不足以说服公众；严重威胁生命安全；空头文件；一纸空文；吃惊；一而再再而三；犯错成本实在太低；重大缺陷；不能用道歉来消除；最强化学致癌物；信心脆弱	S	[–1, –0.8)	37	Y
		A	[–0.8, –0.6)	35	Y
		B	[–0.6, –0.4)	30	N
		C	[–0.4, –0.2)	21	N
		D	[–0.2, 0.1)	15	N

(注：表中"Y"表示"是"，"N"表示"否"。)

由表 2 可知，此三类突发性安全事件的舆情等级皆属于 S 级特重舆情，其中以滨海化工厂爆炸泄露事件最为严重，都需要企业或监管部门高度重视并及时处理。

4　结论与建议

针对目前突发性安全事件预警范围狭窄、透明性不强、反应不够及时等问题，本文将外部性主体考虑在内，进行网络舆情的预警研究。在对热门事件及关键词进行实时挖掘的基础上，结合舆情分类的相关特性，添加权重的基准偏移值，改进 LSA+SVM 算法，通过修正因子的正负值进行舆情倾向性判定及舆情预警分类。

(1) 当修正因子为负且范围在[–1, –0.4)时，根据权重排序确定其为 S、A、B 三个等级，代表该舆情来源负面倾向明显，影响较大，需企业及时介入解决；

(2) 当修正因子范围在[–0.4, 0.1)时，确定该舆情来源属于 C、D 级，代表该舆情属于中性舆情，需要保持观察；

(3) 当修正因子范围在[0.1, 1]时，则判定其为 P 级舆情，属于正面舆情，有助于维持企业积极形象。

利用 LibSVM 和 Matlab 进行仿真和准确率计算，对算法有效性进行了验证，其结果可以体现舆情文本的倾向性，能够为企业提供准确的警报信息。最后，对企业应采取的措施提出以下建议：

S 级(特重舆情)：高度重视和及时应对。企业需要立刻派遣专业的公关团队，迅速锁定舆情源头，及时进行产品召回和赔偿处理，尽可能减少负面影响对企业形象的危害，树立有担当，有责任的企业形象。

A 级(重度舆情)：采取措施解除危机。对于重度舆情需要企业及时介入处理，以免舆论进一步扩散，从而转化为特重舆情。此时企业可以衡量自身资源和危机处理能力，在不损害当前企业利益前提下整合资源，避免危机扩大恶化。

B 级(中度舆情)：抑制舆情信息进一步扩散。严密监测危机信息状态和舆论导向，适当引导舆论；同时启动预案，确保事件向有利方向发展。

C 级(轻度舆情)：排除干扰信息，积极应对。对企

业相关部门提出相应改进意见，持续监测，有变动及时跟进。

D 级(需要关注)：做好日常监测。对于舆情类别进行初判，正面舆情收录到企业创新知识库中，负面舆情作为潜在问题源先行备案，做到防患未然，居安思危。

P 级(信息反馈)：作为反馈建议。由于大多 P 级信息不带过多感情色彩或以正面信息为主，企业可参考反馈信息创新产品、加强管理以及服务升级，为企业发展提供新的思路和契机。

参考文献：

[1] 吴鹏，杨爽，张晶晶，等．突发事件网络舆情中网民群体行为演化的 Agent 建模与仿真研究[J]．现代图书情报技术，2015(7/8): 65-72. (Wu Peng, Yang Shuang, Zhang Jingjing, et al. Agent-Based Modeling and Simulation of Evolution of Netizen Crowd Behavior in Unexpected Events Public Opinion [J]. New Technology of Library and Information Service, 2015 (7/8): 65-72.)

[2] Li W, Miao D, Wang W. Two Level Hierarchical Combination Method for Text Classification [J]. Expert Systems with Applications, 2011, 38(3): 2030-2039.

[3] 王兰成．基于网络舆情分析的突发事件情报支援研究[J]．情报理论与实践，2015，38(7): 72-75. (Wang Lancheng. Research on Emergency Information Support Based on Network Public Opinion Analysis [J]. Information Studies: Theory & Application, 2015, 38(7): 72-75.)

[4] Papetti P, Costa C, Antonucci F, et al. A RFID Web-based Infotracking System for the Artisanal Italian Cheese Quality Trace Ability [J]. Food Control, 2012, 27(1): 234-241.

[5] 董凯欣，傅荧，孙晓峰，等．基于社会网络分析的企业网络舆情预警机制研究——以食品安全网络舆情为例[J]．电子商务，2015, 23(8): 54-55, 57. (Dong Kaixin, Fu Ying, Sun Xiaofeng, et al. Research on Early Warning Mechanism of Enterprise Public Opinion Based on Social Network Analysis [J]. E-Business Journal, 2015, 23(8): 54-55, 57.)

[6] 陈福集，陈婷．舆情突发事件演化探析——基于意见领袖引导作用视角[J]．情报资料工作，2012, 36(2): 23-28. (Chen Fuji, Chen Ting. Research on Public Opinion Emergencies Evolution: Based on the Perspective of Opinion Leaders Guiding Role [J]. Information and Documentation Services, 2012, 36(2): 23-28.)

[7] 宣云干，朱庆华．基于 LSA 的社会化标注系统标签语义检索研究[J]．图书情报工作，2011，55(4): 11-14. (Xuan Yungan, Zhu Qinghua. Research on Tag Semantic Retrieval in Social Tagging System Based on LSA [J]. Library and Information Service, 2011, 55(4): 11-14.)

[8] 范玉华，秦世引．基于潜在语义分析的场景分类优化决策方法[J]．计算机辅助设计与图形学学报，2013，25(2): 175-182. (Fan Yuhua, Qin Shiyin. Optimizing Decision for Scene Classification Based on Latent Semantic Analysis [J]. Journal of Computer-Aided Design & Computer Graphics, 2013, 25(2): 175-182.)

[9] 商丽媛，谭清美．基于支持向量机的突发事件分级研究[J]．管理工程学报，2014，28(1): 119-123. (Shang Liyuan, Tan Qingmei. Emergency Classification Based on Support Vector Machine [J]. Journal of Industrial Engineering and Engineering Management, 2014, 28(1): 119-123.)

[10] 张建娥．基于 TFIDF 和词语关联度的中文关键词提取方法[J]．情报科学，2012，30(10): 1542-1555. (Zhang Jian'e. A Chinese Keywords Extraction Approach Based on TFIDF and Word Correlation [J]. Information Science, 2012, 30(10): 1542-1555.)

[11] 张长利．面向特定领域的互联网舆情分析技术研究[D]．长春：吉林大学，2011. (Zhang Changli. Research on Domain-Oriented Public Sentiment Analysis Technologies [D]. Changchun: Jilin University, 2011.)

[12] 高宏岩，王建辉．在线自调整修正因子模糊控制方法和应用[J]．微计算机信息，2006，22(13): 83-84. (Gao Hongyan, Wang Jianhui. A Fuzzy Control Method with Online Self-turning Correction Factor and Its Application [J]. Microcomputer Information, 2006, 22(13): 83-84.)

[13] Goñi S M, Oddone S, Segura J A. Prediction of Foods Freezing and Thawing Times: Artificial Neural Networks and Genetic Algorithm Approach [J]. Journal of Food Engineering, 2011, 84(1): 164-178.

[14] 谭光兴，刘臻晖．基于 SVM 的局部潜在语义分析算法研究[J]．计算机工程与科学，2016，38(1): 177-182. (Tan Guangxing, Liu Zhenhui. A Local Latent Semantic Analysis Algorithm Based on Support Vector Machine [J]. Computer Engineering and Science, 2016, 38(1): 177-182.)

[15] Sengupta A S, Balaji M S, Krishnan B C. How Customers Cope with Service Failure? A Study Does Brand Reputation and Customer Satisfaction [J]. Journal of Business Research, 2015, 68(3): 655-674.

[16] 朱舸，齐佳音．企业危机事件网络舆情态势评估[J]．情报科学，2015，33(6): 48-53. (Zhu Ge, Qi Jiayin. Situation Evaluation of Online Public Opinion on Enterprise Crisis Event [J]. Information Science, 2015, 33(6): 48-53.)

[17] 马宁，刘怡君．基于超网络的舆情演化多主体建模[J]．系

统管理学报, 2015, 24(6): 785-804. (Ma Ning, Liu Yijun. Multi-Agent Modeling of Public Opinion Evolution Based on SuperNetwork Analysis [J]. Journal of Systems & Management, 2015, 24(6): 785-804.)

作者贡献声明:

田世海: 提出研究思路, 设计研究方案, 论文最终版本修订;
吕德丽: 研究过程实施, 数据获取, 进行实验, 论文撰写。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: lvdelixx@126.com。

[1] 田世海, 吕德丽. pythonurl1.txt.Python 爬虫挖掘代码.
[2] 田世海, 吕德丽. wo3url.csv. 挖取的 URL 列表.
[3] 田世海, 吕德丽. lsasvm.csv. 分词后情感倾向性特征词矩阵.
[4] 田世海, 吕德丽. lsasvmdepart.rdf. 经改进 LSA+SVM 算法计算的特征词向量矩阵.
[5] 田世海, 吕德丽. orien.xls. 文档倾向性准确率列表.

收稿日期: 2016-08-29
收修改稿日期: 2016-10-25

An Early Warning Algorithm for Public Opinion of Safety Emergency

Tian Shihai Lyu Deli
(School of Management, Harbin University of Science and Technology, Harbin 150040, China)

Abstract: [**Objective**] This study proposes a new early warning model to track the public sentiment online, aiming to improve transparency and responding speed of the safety emergencies. [**Methods**] We used the modified LSA+SVM algorithm to build an early warning model, which retrieved public opinion data by meta search. [**Results**] We examined the new model with three different incidents, and found it was practical and fast. The precision rate was 85.75% when the semantic dimension was kept at 10. [**Limitations**] This method was more effective for the safety incidents drawing public attention and discussion. [**Conclusions**] The proposed algorithm helps us build an early warning system for public opinion, which provides suggestions to related companies and government organizations.

Keywords: Latent Semantic Analysis(LSA) Support Vector Machine(SVM) Public Opinion Early Warning Emotional Orientation Analysis